



运用毒理基因组学 进行生物学推论

毒

理基因组学实验中微阵列、蛋白质组学、新陈代谢组学的数据多到不计其数。每一个样本的每一种分析方法就有数以千计,乃至数以万计的数据点,数据的复杂性和绝对数量以倍数递增。当统计学家、生物信息学家和生物学家们对这些复杂的数据进行解释时,他们必须确保每一个点的数据正确,并且能够与来自相同或不同类型实验的其它数据进行整合。这种要求精细的做法主要有两大目标:一是识别有毒物质或疾病的标志;另一个是了解疾病潜在的生物过程,后者也被称之为生物学推论,即用计算机作数学运算,从毒理基因组数据的高反复过程推断其因果关系。

识别暴露标记主要是设法从微阵列、蛋白质组学、新陈代谢组学等技术所获得的结果中获得易分辨的图谱。图谱以分子指纹为特征,在诊断暴露水平中非常有用,即使研究人员可能并不能确切知道为什么会出现独特的图谱。相反,生物学推论所关注的是了解基因组数据转化为基因转录、产生蛋白质和新陈代谢的方式。

通过对基因表达、蛋白质和代谢物相互关系的研究,研究人员试图识别有影响的基因,其中许多基因担当着代谢网络中心的角色,它们影响着其它许多基因。但是很难发现和分析那些在细胞变迁过程中发挥短暂功能的临时网络中心。毒理基因组学的一个极其重要目的是把外源性通道(即由药物或毒物暴露而触发)从内源性通道(与细胞日常繁杂的代谢和繁殖有关)中区分出来。毒理基因组学最终目的是探索通道,了解蛋白质和代谢物的产生和修复中的基因表达,研究所有相关的基因-基因、蛋白质-蛋白质和代谢物-代谢物间的相互作用和交互作用。

分组管理 (Anchor Management)

推断生物学通路需要研究小组挖掘和分析大量的基因组数据,其分析策略包括把数据分组和归类、发现图谱以及利用统计学方法过滤有强信号和表达一致的基因数据。生物学推论的关键技术是从基因组学实验得到的表型簇,即利用已知的生物学信息来解释信号或未确认数据。信号显示了分子(如mRNA或一定分子量的蛋白质)的存在,根据采用的技术不同,信号可以有各种各样的形式。例如,在微阵列实验中,信号以一股mRNA结合在微阵列芯片产生荧光的形式表现出来。在一些研究中,对同一个标本,采用基因组数据与传统的毒物学实验相比较;在另外一些研究中,研究小组根据以往的研究成果,综合已知和尚未确定的生物学通道综合进行分析。

利用表型簇的实例可以在由美国环境卫生研究院(NIEHS)国家毒理基因组中心(NCT)资助的对乙酰氨基酚研究中找到。在这个研究中,微阵列数据与传统毒物学实验数据作了比较,两者均对解释微阵列数据有所帮助,同时对乙

酰氨基酚这个常用药物的毒性有了新的认识和进一步的了解。

例如,发表在《毒理科学》(Toxicological Sciences)2004年7月的一项研究证明,微阵列和传统毒物学实验的数据证实了其它实验室以前的研究结果:即对乙酰氨基酚的有毒剂量耗尽三磷酸腺苷(ATP;储存细胞能量的分子),损伤线粒体和产生ATP的细胞器。另外,微阵列数据显示了其它一些传统实验没有显示的暴露

度遗传上有联系的区域。研究人员对500个基因与QTL区域进行了比较,他们发现一组28个基因表达呈现不同,并且与观察到的表型有连锁遗传关系。用这两种方法识别的基因没有交迭。

一篇发表在2004年6月《基因组学》(Genomics)上的研究报告,研究人员承认采用过滤方法可能遗漏了一些重要的基因。但是,他们认为他们的方法为未来的研究提供了一条得到少量高度优先基因的客观途径。

达、稳定性以及转录片断(多股的RNA)。Affymetrix公司负责生物研究的副总裁Thomas Gingeras说:“要获得基因型-表型的相关性,你需要有每一点转录片断表达的完全目录,在目录中,可以确定基因型-表型的相关性。”

Affymetrix和国家癌症研究所的Gingeras和其他研究人员对10个人类染色体,通过包括全部不重复序列探针(不仅是编码区的转录体)的RNAome微阵列进行了研究。这些阵列数据显示



效应。例如,当乙酰氨基酚剂量非常低,还不足以引起组织病理学和其它传统方法能够发现的细胞损伤时,肝细胞基因表达起始与细胞能量丢失相一致。

微阵列数据也提供一个新的可能是对乙酰氨基酚毒性的信号,包括金属硫蛋白基因和其它几个基因,它们可能与肝脏的抗氧化防御系统有关。NCT科学家、上述论文的主要作者Alexandra Heinloth说:“我们以前不知道这些基因与对乙酰氨基酚毒性有关,但这个发现与生物学上关于细胞防御机制的理论相吻合。”

在炎症相关通道的研究中运用了另一种不同类型的表型簇。通过对小鼠品系吸入脂多糖化合物(触发免疫应答)后高反应水平和低反应水平的微阵列实验,识别了至少在一个品系中起作用的大约500个基因。杜克大学、基因组研究所以及乔治·华盛顿大学的研究人员以及Dana Farber癌症研究所和哈佛大学公共卫生学院生物统计和计算生物学教授John Quackenbush,采用了两种独立的方法过滤微阵列实验结果,把优先基因区分出来以便将来研究。

在第一种方法中,研究小组识别了30个基因,它们的表达水平能够最好的区分高反应和低反应的老鼠。在第二种方法中,他们采用定量特征位点(又称数量性状基因座)(quantitative trait locus, QTL)分析法来发现与脂多糖诱导反应强

人们在许多阵列间能够发现提示共同调节的图谱。如果你观察数以百计的阵列并且发现2个基因的表达同时向上或向下移动,这种现象强烈提示交互作用存在。与其说是生物学模型,还不如说是相互联系的发现提示生物学交互作用存在。

Terry Speed

加利福尼亚大学伯克利分校

微阵列的延伸

加州大学伯克利分校统计学教授Terry Speed说,虽然研究人员的最终目标是将从基因表达到新陈代谢的通道连接起来,因此基因组学研究在很大程度上集中在微阵列,因为这种技术比蛋白质分析(主要是质谱测定)和代谢物分析(质谱测定和核磁共振显微镜)更标准化,并且使用也更加普遍。虽然对蛋白质类似阵列的实验方法已经有了发展,有些利用抗体作为标志,但是这些技术还相对处于探索阶段,而且适用范围有限。

但是用微阵列数据进行生物推断有很大的局限性。虽然微阵列数据可以显示相关性,但是这种数据很少能够提示因果关系。Speed说:“人们在许多阵列中能够发现提示共同调节的图谱。如果你观察数以百计的阵列并且发现2个基因的表达同时向上或向下移动,这种现象强烈提示交互作用存在。与其说是生物学模型,还不如说是相互联系的发现提示生物学交互作用存在。”微阵列数据最主要的局限性之一反映了生物学根本原理:因为静止RNA和其它一些机制能够阻断转译过程,mRNAs的表达不是始终能够转译为蛋白质。

将基因表达和蛋白质制造之间的鸿沟填平的一种方法是采用蛋白质组学分析法对细胞内蛋白质进行评估。另外一种方法是更好地了解“转录体(transcriptome)”(也称为“RNAome”)——即所有调节基因成份的表达,操纵细胞内的调控表

RNA活动性各异、非常复杂。虽然研究人员已经能够识别许多序列的作用,比如核糖体RNA和蛋白质-编码RNA,但是他们不得不对大量的、新发现的转录片断进行了分类,如未知功能转录片断(transcripts of unknown function, TUFs)。

TUFs的序列同样复杂。Gingeras说:“十分令人惊奇,这些转录片断大部分位于基因的中间,交迭在编码序列的义链和反义链上。”义链,或模板链、单链DNA是复制或转录的产物。作者推测与反义链相关的RNA可能是cRNA的拷贝,其产生的方式与在微阵列中采用的cDNA复制RNA的方式相同。

对所有类型的基因组学实验,另外一个挑战是在低水平分子表达中探测信号。样本中mRNA和蛋白质的数量可以相差到1百万比1。这些低表达分子可能是生物学级联效应的重要触发因素,但是含量低的信号可能在信噪比(signal-to-noise)中丢失。

蛋白质组学技术在探测低表达分子中取得了进展,通过更复杂的分类技术,诸如SELDI-TOF(在样本中只选择一个蛋白质子集作分析)。为了在一个样本中发现低表达转录片断和mRNA的数量,研究人员从研究基因表达转向了其它方法,如RT-PCR(它能够利用对照和荧光标记,对聚合酶联反应中产生的分子数量进行定量)和SAGE(它使每一个转录片断有一独特的标记,然

后链接和排序结合的转录片断，以计数每一个标记发生的次数）。

数据分析

就像发展探测和识别转录片断、蛋白质和代谢物技术所面临的挑战一样，分析结果数据的困难可能更大。当研究小组计划他们的实验时，他们必须从各种令人迷惑、不断变化的数据分析和统计方法中作出选择。Speed说：“没有一个方案能解决所有的实验数据。通常情况下，有一个方法是首选的，常常还有几个可以接受的方法。所有方法各有利弊。”

分析结果数据的部分难点与目前基因组学实验有关。传统的统计学方法是基于假设，其研究的样本要比数据点的样本要大得多，而基因组学实验的情况通常正好相反；少量的样本，每个样本有数万个数据点。处理基因组学数据特性的新统计学方法正在发展中。同时，其他研究小组正在努力工作，通过找出实验技术和实验室程序的差异，以及重新评价样本量的影响，以确保分析数据有效。

最近的研究表明，微阵列平台标准化的增加极大地降低了平台类型对结果的影响。发表在2005年5月的《自然法》(Nature Methods)中的一个研究表明，Quackenbush和他的同事们发现微阵列平台类型的不同（寡核苷酸对点状式cDNA）确实会影响结果。然而，不同暴露剂量血

必需看到，研究方案的变化（RNA标记、杂交、和微阵列处理）、统计方法中数据获得和正态化采用的方法以及其它“实验室因素”仍然能显著地影响微阵列数据。这是其它2个研究的发现，也发表在2005年5月出版的《自然法》上，其中一个研究由约翰·霍普金斯大学生物统计系副教授Rafael Irizarry领导；另外一个研究由NIEHS毒理基因组学研究联盟（Toxicogenomics Research Consortium, TRC）的研究人员完成。这两个研究均比较了同一标本多个实验室执行的微阵列分析结果，通过仔细比较，两个研究均发现了不同实验室的试验结果具有可比性。但是，TRC研究的作者之一，华盛顿大学-NIEHS生态地理学和环境健康中心生物信息学和生物统计学部主任Katherine Kerr说：“必须非常小心实验操作的过程，并对各个实验室的研究方案进行标准化”。

微阵列实验设计要求每一个治疗组有3~5个样本。但是，有些统计学家认为必须增加样本量以得到统计学上推断生物学通路所必需的把握度。位于缅因州Bar Harbor市的Jackson实验室的科学家Gary Churchill说：“目前大多数系统生物学实验的样本量不足以推断各种复杂的网络，而这本是研究的目的”。

Churchill是Collaborative Cross的共同创始人之一，Collaborative Cross是一项发展一组1000只新的、遗传多样性小鼠品系的计划。这些老鼠只是8个品系小鼠的后代，从而使得基因型最小

正确地解释原始信号。对于微阵列数据，包括对每一个点的荧光数据（由与探针链接的单个mRNA产生）归纳总结为一个基因产生的一个信号值。Irizarry说：“最高难度是将背景杂信产生的强信号成份剔除”。一些背景杂信（能够与观察到的真实信号混淆的外来信号）能够使得mRNA向芯片附着时发生错配。

Irizarry说另外一种潜在的情况是一些寡核苷酸芯片中的25个碱基对探针可能比其它探针更有“黏附性”——也就是说，它们更能够吸引mRNA。他说：“如果有一个基因被一段具有粘性的序列表现出，那么需要收集一个以上黏附性小的探针”。作为结果，从这样的探针发现的结果可能反映了cDNA-mRNA结合体的化学特性而不是样本的生物学特性。

一旦确定了阵列中的每一个基因，各个阵列中的信号需要标准化。均衡双色芯片中的荧光信号是最基本的标准化方法。如果2个样本杂交到芯片上的数量相等，那么每一个样本上总的荧光信号也应该相同。如果一个样本中的荧光信号呈现均一地增高，统计学家能够对荧光值进行调整，以更好地描述样本之间的关联。一些更复杂的方法可能也用于数据的标准化。

根据TRC研究，标准化过程似乎增加了微阵列数据的精确性。但是关于标准化的公式以及分析基因组学数据的运算法则，还有许多尚未解



目

前大多数系统生物学实验的样本量不足以推断各种复杂的网络，而这本是研究的目的。

统

计学不是与公式和数字打交道，而是概念。它涉及不确定性的量化、数据的收集，并将数据转化为知识。

Gary Churchill
杰克逊实验室

管紧缩素II（引起血管收缩的一种强缩氨酸）引起的平台之间表达的差异与剂量高度相关，这种差异使得不同类型微阵列平台间的差异黯然失色。Quackenbush说：“我们想知道的是：是生物学还是平台占据优势？在我们能够作出合理比较的多于90%的基因中，我们发现生物学比平台类型更重要。”

化，并且将通过繁殖使得基因变化最大化，以获得多样且能控制的品系。Churchill说：“虽然有些研究需要，但是不是所有的研究都需要1000只老鼠”。Collaborative Cross将能满足广泛的需求。

译码的关键

在研究小组开始分析数据前，他们必须确信

决的问题。Kerr说：“当我们采用了‘标准化’，数据看起来比较好。但是，我们不知道我们是否真正作了修正。”

标准化后，研究人员可以对基因表达变化进行基本计数。Churchill称这个过程为“制作数据清单。”他解释：“你测量了正常组织和病变组织的微阵列数据，并且得到了上调或下调的基因数

据清单。问题在于我们怎样把这些数据清单转变为生物学意义。”

数据转化为推论

对于许多研究小组下一步的工作是缩短这个数据清单。他们可能集中关注在那些表达最强或变化最相关的基因。通常这个过程需要对已知基因功能信息全面广泛的了解。这一过程需要格外小心，设置滤器太严格可能引起重要基因被忽略，



如果参数设置范围太宽泛则可能造成假阳性。

宝洁（Procter & Gamble）公司的产品安全方面研究人员Greg Carr说：这就是为什么微阵列实验中，不同研究小组难以复制试验结果。如果探测基因是否存在统计学显著性差异的把握度或概率设置非常低，比如说10%，那么一个实验室可能选取了一些影响力低的基因，而另外一个实验室可能选取其它一些基因，但是，没有一个实验室可能探测到所有的基因。

有时候采用或者代替滤器的另外一种方法是依照表达图谱类似性对数据进行聚类或分组。方法包括分等级、“Eisen”聚类（这种方法将得到一个树状结构）、 κ -均数聚类（这种方法将得到几个线图）和主成分分析（这种方法将得到一个能够旋转的三维排列）。Speed说聚类分析方法通常是探索性的，并且不能对设计好的问题提供答案；它们只能显示基因间的相互联系。但是，它们为组织数据提供了有效的途径，科学家们因此可以推断原因和结果。

各种各样的运算模型能给推论过程提供帮助。最简单的模型之——布尔网络（Boolean networks），能从数据测量推断多变量的基因关系。位于西雅图的系统生物学研究所（Institute for Systems Biology）副教授Ilya Shmulevich，一直致力于通过拓展概率布尔网络来增加布尔网络模型的适用性。Shmulevich说这些网络允许每一个基因有多种多样的功能可能性，模拟潜在的生物学

和测量的不确定性。Shmulevich和他的同事已经运用概率布尔数学体系网络分析了恶性黑素瘤和神经胶质瘤研究的基因表达数据。

运用大多数的模型方法，Speed说：“你不得不只选取非常有限的一些基因进入模型，然后了解这些基因的一些情况”。选取哪些基因进入模型以及决定如何筛选其它基因组学数据，这是一门艺术，对研究人员已有知识的系统了解和掌握有特别高的要求。研究人员梳理了基因组学、蛋

生物学家可能也需要通过培训对统计学有更多的了解。Churchill说：“要深入理解你所需要的概念，选修统计学初级课程并不一定能学到你需要掌握的概念。统计学不是与公式和数字打交道，而是概念。它涉及不确定性的量化、数据的收集，并将数据转化为原理。”

Heinloth说：“如果你在开始这一领域研究以前已经有过统计学的培训那是很好的，但是如果你的研究小组中有足够的统计学家和生物信息

我们需要从探索和理解简单的系统开始，以简化方式去理解复杂问题。如果你从一开始就试图研究复杂问题，你将永远不能真正地把错综复杂的数据梳理清晰。

Kenneth Ramos
路易斯维尔大学

白质组学和新陈代谢组学的科学文献或数据库，但迄今为止数据的挖掘只能做到这一步。NCT领导人Raymond Tennant说只有60%~65%的人类基因的功能得到充分的诠释，这使得推断那些功能没有诠释的基因非常困难。

生物学系统知识的化学效应库（Chemical Effects in Biological Systems Knowledgebase）正在加紧筹建中，计划在2005年末对公众开放。这个数据库将提供约140种参考化合物，以及包括乙酰氨基酚在内的大约10种肝毒物的全面的微阵列数据资料。NCT数据库发展主任助理Michael Waters说：“数据库也将包括化学品和其它物质的生物学效应参考信息，以及与作用机制相关的通路”。

关于推论语言

除了掌握必要的数据，生物学推论需要广泛的技能和专业技术。路易斯维尔大学（University of Louisville）生物化学和分子生物学系主任、EHP毒理基因组学编辑Kenneth Ramos说：“需要有跨学科的知识——注意，我说的是跨学科而不是各学科内部间。我们需要能够说一种以上语言的新型科学家。我们太专业化了，很难在学科间自由流动。我想生物学推论研究需要具备这种能力。传统方法培训出来的生物学家需要在数学方面重新培训以解决生物推论过程中遇到的一些问题。”

学家作为平等的研究成员，那么你已经解决了统计方面的问题”。在Heinloth的研究小组中，统计学家从实验一开始就参与研究。她说：“研究中几乎没有一个决定是由一个人做出的”。但是，她指出合作的过程不是以“科学委员会”的形式，而是在实验设计和执行过程中，小组成员只在彼此的专业领域发挥作用。

因为研究人员需要钻研庞大的实验数据，他们可能面临现有知识的局限。一个人要完全理解即使是单个细胞中奇特和复杂的新陈代谢是不可能的。研究人员要开发基因组学研究的新技术和新统计工具，以对没有研究过的数据作出推论，他们需要开辟新的途径来突破自身理解力的局限。

Ramos说：“我们需要从探索和理解简单的系统开始，以简化方式去理解复杂问题。如果你从一开始就试图研究复杂问题，你将永远不能真正地把错综复杂的数据梳理清晰”。Ramos补充说：“从简单系统开始，会有许多潜在的错误，但只有以这种方式开始，才能逐渐建立更加复杂的模型。”

Churchill说：“当我们收集的数据越来越多，我们意识到我们理解很有局限性，需要开拓的领域还很多。从某种意义上讲可能令人气馁，但是展现在你面前的是一个美妙的新世界。”

—Kris Freeman

译自 EHP 113: A388—A393 (2005)